**GSDC.**
Global Skill Development Council

12-ARTIFACT TOOLKIT BRIEF | CGAIC CERTIFICATION | PRINTABLE PDF

# The full 12-artifact toolkit brief.

Every artifact in this gallery as a build sheet: stack, dataset, training notes, evaluation harness, deployment checklist. Plus the 9-module syllabus and sample exam.

| 12 | 30+ | 2.5L+ |
|---|---|---|
| BUILD SHEETS | LBD LABS | CERTIFIED PROS |

## Inside the toolkit:

- **12 artifact build sheets**
- **Per-artifact stack + dataset + eval**
- **Capstone selection guide (pick 3)**
- **9 module syllabi · verbatim**
- **30+ Learn-by-Doing labs catalog**

---

**Program:** Certified Generative AI in Cybersecurity (CGAIC)

**Exam:** 40 MCQ · 90 min · free retake | **Duration:** 90 days

Used by 2,50,000+ certified professionals worldwide.

## The 12 Artifacts · Catalog Overview

Every artifact is something you *build* — not read, not summarise. Each ships with a fixed stack, a labelled dataset, training notes, an evaluation harness, and a deployment checklist. The 12 cover the full AI-security work surface across defence, detection, and offence.

| # | Artifact | Category | Stack family |
|---|---|---|---|
| 01 | **GAN-based Network Anomaly Detector** | Detection | PyTorch / GAN |
| 02 | **VAE Behaviour Analytics Engine** | Detection | PyTorch / VAE |
| 03 | **RL-based Incident Responder** | Response | Stable-Baselines3 / RL |
| 04 | **Agentic SOC Triage Pipeline** | Response | LangGraph / LLM agents |
| 05 | **OWASP LLM Top 10 Audit Kit** | Audit | Python · test harness |
| 06 | **Prompt-Injection Test Harness** | Audit | Python · garak / promptmap |
| 07 | **MITRE ATLAS Detection Coverage Heatmap** | Threat Intel | SIEM + spreadsheet |
| 08 | **LLM Guardrail Kit** | Defence | Python · guardrails library |
| 09 | **Hardened RAG Pipeline** | Defence | LangChain / Llama-Index |
| 10 | **AI-Phishing Detection Classifier** | Detection | scikit-learn / fine-tuned LLM |
| 11 | **Red-Team Engagement Reporting Kit** | Offence | Markdown · Burp · custom |
| 12 | **AI-HR Governance & Risk Memo** | Governance | Word · NIST/ISO refs |

### How the catalog is structured

- **5 categories.** Detection (3) · Response (2) · Audit (2) · Threat Intel (1) · Defence (2) · Offence (1) · Governance (1).
- **Each build sheet has 5 sections.** Stack · Dataset · Training notes · Evaluation harness · Deployment checklist.
- **Every artifact is interview-ready.** You walk an evaluator through it once at capstone defence; you walk hiring managers through it for the next two years.
- **The 12 are also the capstone shortlist.** You pick 3 of these 12 for your CGAIC capstone (selection guide on page 16).

> *The artifacts are the credential. The badge gets you the screen; the artifacts get you the offer.*

## Build Sheets · Artifacts 1–2 · Detection (Generative)

### 1 GAN-based Network Anomaly Detector

**DETECTION · 8–12 HOURS · PYTORCH**

A Generative Adversarial Network that learns the normal distribution of network-flow features and flags samples the discriminator scores as **out-of-distribution**. Catches novel attacks that signature-based tools miss.

**STACK**
PyTorch, NumPy, scikit-learn, pandas, MLflow for run tracking.

**DATASET**
CIC-IDS2017 or UNSW-NB15, plus a held-out synthetic novel-attack split.

**TRAINING NOTES**
Generator + discriminator, ~150 epochs, batch 256, Adam (β1 = 0.5). Mode-collapse the main failure.

**EVALUATION**
AUC-PR on benign vs novel-attack, per-attack-family confusion, drift KPI over weekly slices.

**DEPLOY** Wrap as gRPC service · model card with intended use + limits · monitoring on score-distribution drift · monthly retrain trigger.

### 2 VAE Behaviour Analytics Engine

**DETECTION · 8–12 HOURS · PYTORCH / VAE**

A Variational Autoencoder over user/entity behaviour sequences. Reconstruction error scores per session; sustained high error = anomalous behaviour. Strong on **insider-threat & account-takeover** patterns where signatures fail.

**STACK**
PyTorch Lightning, pandas, FastAPI for inference, Postgres for session features.

**DATASET**
CERT Insider Threat dataset (r4.2 or r6.2) plus your own sanitised auth/activity logs.

**TRAINING NOTES**
LSTM-VAE on 50-event windows. KL-anneal over 20 epochs. Per-user baseline subtraction.

**EVALUATION**
Reconstruction-error threshold tuning on per-user holdout. Precision @ top-1% of sessions is the headline KPI.

**DEPLOY** Daily batch scoring · per-user baseline auto-refresh · SOC console integration · false-positive feedback loop.

**Why these two artifacts first:** They demonstrate *generative model fluency* on a security problem — the strongest single signal for the AI Security Engineer hiring loop. Most candidates without certification can describe these architectures; few can show working examples with labelled evaluation.

## Build Sheets · Artifacts 3–4 · Response & Agentic

### 3 RL-based Incident Responder

**RESPONSE · 10–14 HOURS · STABLE-BASELINES3 / RL**

A Deep Reinforcement Learning agent (PPO or DQN) that learns to recommend **containment actions** against a simulated network — isolate host, block IP, kill process, escalate. Trained in a CybORG-style gym environment.

**STACK**
Stable-Baselines3, Gymnasium, CybORG (CAGE Challenge), PyTorch, Weights & Biases.

**DATASET**
Simulated incident trajectories — no live data. Curriculum from CAGE Challenge red/blue scenarios.

**TRAINING NOTES**
PPO with shaped reward (containment +, business impact −). 1–3M timesteps. Reward hacking is the main risk.

**EVALUATION**
Mean episodic return on held-out scenario set; head-to-head vs a scripted-rule baseline. Catastrophic-action rate must be ~0.

**DEPLOY** Recommendation-only mode for first 30 days · human-in-the-loop approval · audit log of every suggested action · rollback runbook.

### 4 Agentic SOC Triage Pipeline

**RESPONSE · 10–14 HOURS · LANGGRAPH / LLM AGENTS**

A multi-agent LangGraph pipeline that takes a raw SIEM alert and produces an **enriched triage report** — IOC enrichment, MITRE ATLAS mapping, suggested containment, and confidence score. Replaces the first 15 minutes of Tier-1 analyst work.

**STACK**
LangGraph, LangChain, Anthropic/OpenAI API, VirusTotal & AbuseIPDB connectors, FastAPI gateway.

**DATASET**
200 historical alerts (anonymised) with analyst-graded ground-truth triage; eval split of 40.

**TRAINING NOTES**
No training — pure prompting + tool use. Prompt versioning is the main discipline; eval-driven prompt iteration.

**EVALUATION**
Triage-accuracy vs analyst ground truth (F1 on 4 severity buckets), enrichment-completeness score, time-to-first-recommendation.

**DEPLOY** Co-pilot mode in SOC console · per-step audit log · output-cap to prevent prompt-injection-driven false escalations · weekly prompt-quality review.

## Build Sheets · Artifacts 5–6 · Audit

### 5 OWASP LLM Top 10 Audit Kit

**AUDIT · 6–10 HOURS · PYTHON TEST HARNESS**

A reusable audit kit that walks an LLM application end-to-end against the **OWASP LLM Top 10**. Each category has a test pack, a severity matrix, and a remediation memo template. Output is an 8–12 page audit report.

**STACK**
Python, pytest-style harness, OpenAPI for target apps, Jinja2 for report rendering, CSV for severity matrix.

**DATASET**
Curated test-prompt library (~200 prompts) covering all 10 OWASP categories, with expected-behaviour labels.

**TRAINING NOTES**
No training — assertion-based testing. The discipline is in writing tight pass/fail criteria per category.

**EVALUATION**
Pass-rate per OWASP category, severity-weighted score, false-positive review of any failed assertion before reporting.

**DEPLOY** CI integration on every release · audit-report archive · severity-tier-to-blocker mapping · re-test schedule per remediation.

### 6 Prompt-Injection Test Harness

**AUDIT · 6–10 HOURS · PYTHON · GARAK / PROMPTMAP**

A targeted harness that runs **direct + indirect prompt-injection attacks** against an LLM application — encoded payloads, multi-turn jailbreaks, RAG-poisoning vectors, tool-call hijack. Output is a reproducible PoC repo + remediation memo.

**STACK**
garak, promptmap, custom payload generators, pytest, requests for the target API.

**DATASET**
~500 attack prompts across 6 OWASP LLM01 sub-classes, plus 50 RAG-poisoning seeds.

**TRAINING NOTES**
No training. Curate the prompt library on real-world recent jailbreaks; refresh quarterly.

**EVALUATION**
Bypass-rate per sub-class, time-to-first-bypass, mean attack length until success. Goal is full coverage, not max bypass.

**DEPLOY** Run before every prompt-template change · pair with input/output filter (Artifact 8) regression suite · publish a fixed-version baseline before any "we are protected" claim to leadership.

# Build Sheets · Artifacts 7–8 · Threat Intel + Defence

## 7 MITRE ATLAS Detection Coverage Heatmap

**THREAT INTEL · 8–12 HOURS · SIEM + SPREADSHEET**

A heatmap of tactic-vs-data-source coverage for AI attacks, plus **5 SIEM-portable detection rules**. The artefact your SOC Lead points at when asked "what AI attacks can we actually catch?"

**STACK**
Splunk / Sentinel / Elastic detection-as-code, MITRE ATLAS Navigator, spreadsheet for coverage matrix.

**DATASET**
Map your existing data sources (proxy, EDR, model gateway, vector DB audit) against ATLAS tactics relevant to your stack.

**TRAINING NOTES**
No training. The work is in writing detection rules that fire on real signals, not synthetic ones.

**EVALUATION**
Per-rule true-positive count over a 30-day live window, false-positive rate, time-to-tune. Document "intentional gaps" as a feature.

**DEPLOY** Detection-as-code repo · quarterly ATLAS-version sync · ATLAS → ATT&CK bridge table · monthly purple-team validation.

## 8 LLM Guardrail Kit

**DEFENCE · 8–12 HOURS · PYTHON GUARDRAIL LIBRARY**

An installable Python library that wraps an LLM call with **input filters, output validators, scope-limiting agents, and tamper-evident audit logging**. Bench-tested against the OWASP LLM Top 10 and the Prompt-Injection harness (Artifact 6).

**STACK**
Python, Guardrails AI or NeMo Guardrails, custom regex/embedding filters, pytest for regression.

**DATASET**
Filter test corpus from Artifact 6 plus an in-distribution allow-set so you can measure false rejection.

**TRAINING NOTES**
Embedding-based filters need a calibration step on benign in-distribution traffic; refresh quarterly.

**EVALUATION**
Block-rate on Artifact 6 corpus, false-reject rate on benign queries, latency overhead per request.

**DEPLOY** Wrap every LLM gateway · feature-flag rollout · monitoring on block-rate and latency budget · runbook for "filter is blocking a legitimate request" tickets.

# Build Sheets · Artifacts 9–10 · Defence + Phishing

## 9 Hardened RAG Pipeline

**DEFENCE · 10–14 HOURS · LANGCHAIN / LLAMA-INDEX**

A reference RAG implementation with end-to-end controls — **source vetting, per-tenant scoping, content sanitisation, retrieval-result filtering, response validation, tamper-evident logging**. The architecture you'd hand to engineering to copy.

**STACK**
LangChain or Llama-Index, Pinecone or Weaviate, OPA for policy, Guardrails AI for output validation.

**DATASET**
Three test corpora — benign documents, poisoned documents (Artifact 6 RAG seeds), cross-tenant leak attempts.

**TRAINING NOTES**
No model training. Embedding-model selection is the main lever; baseline against text-embedding-3 and a domain-tuned variant.

**EVALUATION**
Cross-tenant leak rate (must be 0), poisoning bypass rate, response-validity rate, end-to-end latency.

**DEPLOY** Threat-model walkthrough as a deploy gate · per-tenant smoke tests in CI · vector-DB audit log retention · quarterly external review.

## 10 AI-Phishing Detection Classifier

**DETECTION · 6–10 HOURS · SCIKIT-LEARN / FINE-TUNED LLM**

A binary classifier that scores incoming email + URL features for **AI-generated phishing** likelihood. Hybrid: a fast lexical/feature model + an LLM second-pass for ambiguous cases. Built for SOC integration, not endpoint use.

**STACK**
scikit-learn (XGBoost), optional fine-tuned small LLM (Llama-3 8B or similar), FastAPI inference, Postgres feature store.

**DATASET**
SpamAssassin + Nazario phishing corpora, plus a synthetic AI-generated phishing set; held-out test of recent real-world samples.

**TRAINING NOTES**
Stratified split by year so you train on past and test on recent. Calibrate probability with Platt scaling for SOC-friendly outputs.

**EVALUATION**
AUC-ROC, AUC-PR (heavily imbalanced), per-month performance over the last 6 months. Cost-weighted error matters more than raw accuracy.

**DEPLOY** Inline in mail gateway in shadow mode for 30 days · per-segment performance dashboards · monthly drift review · feedback loop from SOC verdicts.

## Build Sheets · Artifacts 11–12 · Offence + Governance

### 11 Red-Team Engagement Reporting Kit

**OFFENCE · 8–12 HOURS · MARKDOWN · BURP · CUSTOM**

A complete **AI red-team engagement report** against a target — scope, methodology, exploit chain demo, findings, severity matrix, remediation memo, executive summary. OWASP-LLM-Top-10-aligned. The deliverable hiring managers walk through in interviews.

**STACK**
Markdown for report, Git repo for exploit chain, Burp Suite or custom scripts, screen recording, OWASP LLM Top 10 reference.

**DATASET**
Findings & PoCs from a sanctioned target (lab environment or your own employer with permission).

**TRAINING NOTES**
No training — process and report-writing discipline. The methodology section is what experienced red-team managers grade hardest.

**EVALUATION**
Coverage breadth (OWASP categories touched), depth per finding (PoC + impact narrative), remediation quality, exec-summary readability.

**DEPLOY** Engagement intake template · responsible-disclosure handling rules · remediation tracking integration with engineering board · post-engagement debrief.

### 12 AI Governance & Risk Memo

**GOVERNANCE · 5–8 HOURS · WORD · NIST/ISO REFS**

A 6–10 page governance memo mapping **NIST AI RMF + EU AI Act + ISO/IEC 42001** obligations to your AI-tooling stack. Includes a risk-tier classification per use case, a control matrix, and a 12-month roadmap.

**STACK**
Word / Google Docs, NIST AI RMF + ISO 42001 reference tables, existing security policy library, legal-review handshake template.

**DATASET**
Your enterprise's top 8 AI use-cases (or representative ones if you're building this for a portfolio).

**TRAINING NOTES**
No training — research and writing. The discipline is matching language to your existing security policy register so the memo plugs into the broader programme.

**EVALUATION**
Coverage of NIST RMF functions, EU AI Act risk-tier accuracy, ISO 42001 clause traceability, exec-summary clarity (CHRO/CISO readable in 5 minutes).

**DEPLOY** Quarterly refresh cycle · new-use-case intake gate · regulator-facing summary version · evidence index linking each claim to a control artefact.

## Per-Artifact Stack Summary

All 12 artifacts on a single page, by tool category. Use this to plan tool access before you enrol. Most candidates already have ~60% of the stack at work.

| # | Artifact | ML lib | LLM API | Pipeline / Infra | Eval & Repro |
|---|---|---|---|---|---|
| 01 | **GAN Anomaly Detector** | **PyTorch** | — | FastAPI · MLflow | scikit-learn metrics |
| 02 | **VAE Behaviour Analytics** | **PyTorch Lightning** | — | FastAPI · Postgres | Pandas eval split |
| 03 | **RL Incident Responder** | **Stable-Baselines3** | — | Gymnasium · CybORG | W&B run tracking |
| 04 | **Agentic SOC Triage** | — | **Claude / GPT-4o** | LangGraph · FastAPI | Custom eval harness |
| 05 | **OWASP LLM Audit Kit** | — | **Target app's LLM** | pytest harness · Jinja | Assertion library |
| 06 | **Prompt-Injection Harness** | — | **Target app's LLM** | garak · promptmap | Pass/fail tables |
| 07 | **MITRE ATLAS Heatmap** | — | — | SIEM as code | Detection backtest |
| 08 | **LLM Guardrail Kit** | — | **Guardrails AI / NeMo** | Python package · pytest | Block-rate dashboard |
| 09 | **Hardened RAG Pipeline** | — | **LangChain / LI** | Pinecone · OPA | 3-corpus eval |
| 10 | **AI-Phishing Classifier** | **XGBoost · LLM FT** | **Optional** | FastAPI · Postgres | Cost-weighted matrix |
| 11 | **Red-Team Report Kit** | — | — | Markdown · Git · Burp | OWASP severity matrix |
| 12 | **Governance & Risk Memo** | — | — | Word · ref libraries | Clause traceability |

### Minimum personal stack to ship all 12

- **Python 3.11+** with PyTorch, scikit-learn, pandas. A small workstation works; nothing here needs a multi-GPU rig.
- **One frontier LLM API** with budget for ~20K calls (OpenAI, Anthropic, or Google). The lab kits ship cost-control wrappers.
- **One SIEM access** (your employer's, or a free Elastic / Wazuh stack for the lab work on Artifact 7).
- **Git + a notebook environment** (VS Code or Cursor). No additional licensed tooling required.

**What you don't need:** a commercial AI red-team subscription, an enterprise SIEM licence, or a vendor-specific certification. Every artifact ships on a personal stack — you can build them outside your employer's tooling if you need to.

## 9-Module CGAIC Syllabus (Verbatim)

All 9 modules of the Certified Generative AI in Cybersecurity program. Each artifact in the catalog is built inside one or more of these modules; the mapping is on page 11.

MODULE 01
### Foundations · LLMs for Security Pros
How LLMs work end-to-end at the depth a security professional needs. Tokenisation, attention, RAG, agents, tool calls.

MODULE 02
### AI Threat Landscape
MITRE ATLAS taxonomy, OWASP LLM Top-10, attacker motivations, AI-specific kill chain. Maps to traditional MITRE ATT&CK.

MODULE 03
### Gen-AI Phishing & Social Engineering
AI-generated phishing, deepfake voice/video, BEC variants, detection signatures, user-side defences.

MODULE 04
### AI-Augmented Malware
Polymorphic payloads, AI-generated obfuscation, prompt-injection-based C2, defender techniques.

MODULE 05
### Prompt Injection & LLM Exploitation
Direct + indirect injection, jailbreak chains, model extraction, training-data leakage, embedding attacks.

MODULE 06
### Secure-by-Design for AI Systems
Guardrails, input/output filters, scope-limiting agents, threat modelling for AI features.

MODULE 07
### MLOps Security & Supply Chain
Model registry, signing & provenance, supply-chain attacks, monitoring, rollback, secret scanning.

MODULE 08
### AI Governance, Risk & Compliance
NIST AI RMF, ISO/IEC 42001, EU AI Act, NYC LL 144, board reporting, vendor governance.

MODULE 09
### Capstone · Defend & Certify
Pick 3 artifacts from the catalog, build & defend them in front of an evaluator, earn the CGAIC credential.

Total program time: 90 days, 6–8 hours per week. Exam format: 40 MCQ in 90 minutes with a free retake on first failure.

## Module-to-Artifact Mapping

Which CGAIC module ships which artifact, at what depth. Use this to plan your build sequence if you're prioritising specific artifacts for your capstone.

| Module | Primary artifacts shipped | Supporting |
|---|---|---|
| M01 · Foundations · LLMs | — | Pre-reqs for all LLM artifacts (4, 5, 6, 8, 9) |
| M02 · Threat Landscape | 07 ATLAS Heatmap | Feeds into 11 Red-Team Report |
| M03 · GenAI Phishing | 10 AI-Phishing Classifier | Supports 04 Agentic Triage |
| M04 · AI-Augmented Malware | 01 GAN Anomaly Detector | 02 VAE Behaviour Analytics |
| M05 · Prompt Injection & LLM Exploitation | 06 Prompt-Injection Harness · 11 Red-Team Report | Pairs with 08 Guardrails |
| M06 · Secure-by-Design | 08 LLM Guardrail Kit · 09 Hardened RAG | 05 OWASP Audit Kit |
| M07 · MLOps Security | 02 VAE · 03 RL Responder (training-pipeline rigour) | Hardens deploy of 01, 10 |
| M08 · Governance & Compliance | 12 Governance & Risk Memo | References for all artifacts' model cards |
| M09 · Capstone · Defend & Certify | Defence of any 3 artifacts | — |

### Quick reference · artifacts by category

- **Detection (3):** 01 GAN, 02 VAE, 10 AI-Phishing — generative + classification models for network, user, and email channels.
- **Response (2):** 03 RL Responder, 04 Agentic SOC Triage — the autonomous + co-pilot pair for the SOC.
- **Audit (2):** 05 OWASP Audit Kit, 06 Prompt-Injection Harness — the two assessment tools every AI security team needs.
- **Defence (2):** 08 Guardrail Kit, 09 Hardened RAG — what you ship around production LLM apps.
- **Threat Intel (1):** 07 ATLAS Heatmap — the visibility map.
- **Offence (1):** 11 Red-Team Report Kit — the engagement deliverable.
- **Governance (1):** 12 Governance & Risk Memo — the regulator-facing artefact.

## Capstone Selection Guide · Pick 3 of 12

The CGAIC capstone asks you to defend 3 artifacts from the 12. Pick by your *target role family* in the 90 days after the credential. Four standard combinations cover ~85% of candidate situations.

PICK A · IF YOU'RE TARGETING · AI SOC ANALYST

### Artifacts 04 + 07 + 10

Agentic SOC Triage + ATLAS Coverage Heatmap + AI-Phishing Classifier. The blue-team operator's portfolio — **one autonomous triage agent**, **one coverage map**, **one production classifier**. Hiring managers ask about the heatmap every time.

PICK B · IF YOU'RE TARGETING · GENAI RED TEAMER

### Artifacts 05 + 06 + 11

OWASP LLM Audit Kit + Prompt-Injection Harness + Red-Team Engagement Reporting Kit. The offensive portfolio — **audit framework**, **attack harness**, **full engagement deliverable**. Defend the report; offer the others. The engagement report is the single most-asked-about artefact in red-team interviews.

PICK C · IF YOU'RE TARGETING · AI SECURITY ENGINEER

### Artifacts 08 + 09 + 06

LLM Guardrail Kit + Hardened RAG Pipeline + Prompt-Injection Harness. The platform-engineer portfolio — **two production-grade defences** tested against **one industry-standard attack harness**. The architecture review round will ask about the threat model on the RAG pipeline.

PICK D · IF YOU'RE TARGETING · AI GOVERNANCE LEAD

### Artifacts 12 + 05 + 07

Governance & Risk Memo + OWASP LLM Audit Kit + ATLAS Coverage Heatmap. The CHRO/CISO-track portfolio — **regulator-facing memo**, **repeatable audit framework**, **defensible threat-coverage map**. Used by candidates moving from GRC into AI governance leadership.

### How to think about your pick

- **Cover 2+ categories.** Three detection artifacts looks one-dimensional. Detection + audit + governance looks senior.
- **Pick at least one artifact you've never built before.** The capstone is your chance to add a capability, not just polish existing work.
- **Match to your interview loop.** System-design rounds favour Pick C. Threat-modelling rounds favour Pick B. Compliance rounds favour Pick D.
- **Defend the hardest, offer the rest.** 30-minute defence forces depth on one. The other two get a 5-minute walkthrough.

## 30+ Learn-by-Doing Labs · Catalog (1–16)

Each lab is a time-boxed, evaluator-reviewed exercise tied to one or more artifacts. You finish each lab with a fragment of an artifact — or a complete one, for the shorter labs.

- **01 LLM Tokeniser & Embedding Lab**
- **02 Prompt-Injection Attack Lab (basic)**
- **03 Indirect-Injection via RAG**
- **04 Output-Filter Bypass Bench**
- **05 ATLAS Threat-Model Workshop**
- **06 ATLAS → ATT&CK Bridge Table**
- **07 Detection-Coverage Heatmap Build**
- **08 SIEM Rule · AI Phishing Pattern**
- **09 SIEM Rule · Prompt-Injection C2**
- **10 Deepfake-Voice Detection Triage**
- **11 GAN Anomaly Detector · Train**
- **12 GAN Detector · Eval Harness**
- **13 VAE Behaviour Engine · Train**
- **14 RL Responder · Gym Setup**
- **15 RL Responder · Reward Shaping**
- **16 AI Incident Response Runbook**

### How a lab is structured

- A **2–4 hour** time-box with a clear deliverable.
- A guided **scaffold notebook + dataset + tool stack** you can replicate in your environment.
- **Evaluator review** on output quality plus written feedback to apply on the next lab.
- A reusable **fragment**: a trained model, a SIEM rule, an audit memo, a test harness component — that plugs into one of the 12 artifacts.

## 30+ Learn-by-Doing Labs · Catalog (17–32)

The second half of the labs catalog focuses on architecture, agentic pipelines, advanced red-team, and capstone-track artifacts.

- 17 ASVS L2 Verification Sprint
- 18 Guardrail Kit · Working Code
- 19 Hardened RAG Architecture Build
- 20 Vector-DB Security Audit
- 21 MLOps · Signing & Provenance
- 22 Supply-Chain · LoRA Backdoor Lab
- 23 Model Registry Hardening
- 24 Model Monitoring & Drift Alerts
- 25 MS AI Red Team · Engagement Scope
- 26 Jailbreak Chain · Reproducible PoC
- 27 Red-Team Report · OWASP Format
- 28 Agentic SOC Triage · LangGraph Build
- 29 Vendor Governance Assessment
- 30 EU AI Act · Risk Classification Lab
- 31 Board-Pack One-Pager · AI Risk
- 32 Capstone Build & Defence

### If you only have time for six labs

The minimum portfolio for an AI-security interview loop, in priority order: **Lab 11–12** (GAN train + eval), **Lab 18** (Guardrail kit), **Lab 19** (Hardened RAG), **Lab 27** (Red-team report), **Lab 28** (Agentic SOC), **Lab 32** (Capstone). Together they touch every artifact category at a defensible depth.

### What's not in the lab catalog (and why)

- **Production penetration testing against third-party systems.** Out of scope legally; engagements happen against approved targets only.
- **Vendor-specific tooling certifications.** CGAIC is vendor-neutral by design. You'll touch many tools; you won't earn a tool badge.
- **Pure data-science labs.** Model architecture optimisation, hyperparameter tournaments. CGAIC focuses on securing AI, not building it from scratch.

## Sample Exam — Part 1 of 2

Six representative questions across the 12 artifacts. The real CGAIC exam is 40 MCQ in 90 minutes with a free retake on first failure.

**Q1 · ARTIFACT 1 · GAN ANOMALY DETECTOR**

**Your GAN-based anomaly detector trains stably but converges to a generator that produces only one type of benign sample. The correct diagnosis is:**

(a) Discriminator overfitting; add dropout.

(b) Mode collapse; introduce mini-batch discrimination or use Wasserstein-GAN with gradient penalty.

(c) Learning rate too low; raise it 10×.

(d) Insufficient training data; the architecture is fine.

**Q2 · ARTIFACT 3 · RL INCIDENT RESPONDER**

**Your PPO-trained responder produces high cumulative reward in simulation but takes catastrophic actions (mass isolation) in live shadow mode. The most likely root cause is:**

(a) The discount factor γ is too high.

(b) Reward hacking — your reward function rewards containment without sufficiently penalising business impact.

(c) Insufficient episodes; train for 10× longer.

(d) The simulator is out of date; refresh weekly.

**Q3 · ARTIFACT 5 · OWASP LLM AUDIT KIT**

**When a chatbot's customer-support feature ignores its system prompt when asked in base64-encoded form, the most accurate OWASP LLM Top 10 category is:**

(a) LLM01 · Prompt Injection.

(b) LLM06 · Sensitive Information Disclosure.

(c) LLM08 · Excessive Agency.

(d) LLM10 · Model Theft.

## Sample Exam — Part 2 of 2

**Q4 · ARTIFACT 9 · HARDENED RAG PIPELINE**

**A RAG application retrieves documents from a vector store that any tenant can write to. The single highest-impact mitigation to ship first is:**

(a) Add a profanity filter on the LLM response.

(b) Add a per-tenant retrieval scope so retrieval only returns documents owned by the requesting tenant.

(c) Increase the LLM temperature to add response variety.

(d) Cache responses for 1 hour.

**Q5 · ARTIFACT 7 · ATLAS COVERAGE HEATMAP**

**Your SOC team claims "full ATLAS coverage" but every detection it cites was already firing on classical MITRE ATT&CK techniques. The correct response is:**

(a) Accept the claim; ATLAS and ATT&CK overlap heavily.

(b) Build per-tactic detection traceability and re-evaluate; many SOCs claim ATLAS coverage but only cover the ATT&CK overlap.

(c) Switch SIEM vendors.

(d) Train the SOC team on ATLAS only.

**Q6 · ARTIFACT 12 · GOVERNANCE MEMO**

**Under the EU AI Act, an LLM-based resume screener used in EU hiring is most accurately classified as:**

(a) Minimal risk — no obligations.

(b) Limited risk — transparency obligations only.

(c) High risk — full conformity assessment, registration, and human-oversight obligations.

(d) Prohibited — cannot be deployed in the EU.

### Answer key

**Q1 — b · Q2 — b · Q3 — a · Q4 — b · Q5 — b · Q6 — c**

### If you scored 5–6 of 6

You already think like an AI-security engineer. The capstone defence is where the real differentiation happens — pick the harder Pick B or Pick C combination on page 12.

### If you scored 3–4 of 6

Foundations are solid; you have gaps on the production-engineering side (RL reward shaping, GAN failure modes). The 90-day program is well-paced for you — most candidates in this band land at 90%+ on the real exam.

# Printable Build Checklist · Before You Start an Artifact

Tear this page out or print it. Run through this before opening tooling on any of the 12 artifacts. Five minutes here saves three hours of rework on the build.

## Scope & success criteria

- ☐ You can state the **artifact's purpose in one sentence** without using the word "AI."
- ☐ You know the **single user role** who will read or use the output (SOC analyst, red team manager, CISO, auditor).
- ☐ You know the **"good enough" definition** — what specifically makes this artifact pass evaluator review.
- ☐ You've set a **time-box** (and an alarm). Most artifacts overshoot when un-boxed.

## Data & access

- ☐ Source data identified, with permission to use it, or a public/synthetic substitute.
- ☐ **Data sensitivity classified** — and you're working at the right tooling tier for the classification.
- ☐ Identifiers stripped from anything that leaves your machine or your employer's environment.
- ☐ If using your employer's data, you've informed the right stakeholder before you start.

## Tool stack & evaluation harness

- ☐ All tools in the per-artifact stack table on page 9 are working — not just installed.
- ☐ **Evaluation harness exists before you train**. If you can't measure it, you can't ship it.
- ☐ You have **one baseline** — a simple rule, regex, or scripted policy — to beat. "Better than nothing" isn't an evaluation.
- ☐ You know where the artifact will live after build — your portfolio repo, GitHub, or a private folder.

## Defence & reuse

- ☐ You can name **one trade-off** you'll be asked to defend — and your one-sentence answer.
- ☐ You can name **one failure mode** you've already encountered — and what you did about it.
- ☐ You've written down **one improvement** you'd make in a v2, even before finishing v1.
- ☐ The artifact is reusable for at least two of the four capstone picks on page 12 — or you've decided it's a single-purpose deliverable.

## Done definition

- ☐ Evaluator-graded output is in the format the spec calls for (model + report / harness + score / memo).
- ☐ Time-box honoured — or explicit note of why it was extended.
- ☐ Model card / runbook / README written. The artefact without docs is not finished.
- ☐ Portfolio location updated; link added to your resume or your LinkedIn 'Projects' section.

## Glossary & About This Brief

### Glossary

- **CGAIC:** Certified Generative AI in Cybersecurity — the GSDC vendor-neutral AI-security credential.
- **GAN:** Generative Adversarial Network — a generator + discriminator pair used here for anomaly detection.
- **VAE:** Variational Autoencoder — encoder + decoder + probabilistic latent space, used for behaviour-sequence anomaly scoring.
- **RL · PPO / DQN:** Reinforcement Learning algorithms; PPO (Proximal Policy Optimisation) is the default for the incident-responder artifact.
- **RAG:** Retrieval-Augmented Generation — architecture where an LLM is grounded with retrieved documents at query time.
- **Agentic pipeline:** A multi-step LLM workflow with tool use, memory, and conditional routing (typically LangGraph).
- **OWASP LLM Top 10:** The current OWASP top-10 application-security risks specific to LLM applications.
- **MITRE ATLAS:** The Adversarial Threat Landscape for AI Systems — MITRE's tactics-and-techniques framework for AI attacks.
- **NIST AI RMF:** The U.S. NIST AI Risk Management Framework with four functions — Govern, Map, Measure, Manage.
- **ISO/IEC 42001:** The international standard for an AI Management System (AIMS) — certifiable, like ISO 27001.

### About the Global Skill Development Council

GSDC is a global, independent skill-certification body building worldwide credentials for the future of work. The CGAIC program is part of GSDC's portfolio of AI-era professional certifications — designed with practitioners, validated by mentors actively working in the field, and trusted by 2,50,000+ certified professionals across 45+ countries.

### Verifying your credential

Once you complete the 40-MCQ assessment and the capstone defence on 3 artifacts, your CGAIC credential is issued with a unique verification ID. Recruiters and hiring managers can verify the credential directly on the GSDC registry — no third-party validation needed.

# The 12-Artifact Toolkit Brief · On One Page

## The 12 artifacts (pages 3–8)

GAN Anomaly Detector · VAE Behaviour Analytics · RL Incident Responder · Agentic SOC Triage · OWASP LLM Audit Kit · Prompt-Injection Harness · ATLAS Coverage Heatmap · LLM Guardrail Kit · Hardened RAG Pipeline · AI-Phishing Classifier · Red-Team Engagement Report · Governance & Risk Memo.

## Each build sheet (pages 3–8)

5 sections per artifact: Stack · Dataset · Training notes · Evaluation harness · Deployment checklist. The structure mirrors what a hiring manager will probe in interviews.

## The tool stack (page 9)

Minimum personal stack: Python 3.11+, one frontier LLM API, one SIEM access, Git + a notebook environment. ~60% is already on most candidates' work machines.

## The 9-module CGAIC syllabus (page 10)

Foundations · Threat Landscape · GenAI Phishing · AI-Augmented Malware · Prompt Injection & Exploitation · Secure-by-Design · MLOps Security · Governance · Capstone. 40 MCQ exam · free retake.

## Capstone picks (page 12)

Pick 3 of 12 for your defence. Four standard picks cover 85% of candidate situations: SOC Analyst (4+7+10) · Red Teamer (5+6+11) · Security Engineer (8+9+6) · Governance Lead (12+5+7).

## Labs catalog (pages 13–14)

30+ labs grouped into foundations, threat-modelling, detection-engineering, MLOps, agentic, red-team, governance, and capstone. Minimum-viable lab portfolio: 6 labs covering every artifact category.